Moments of a Probability Distribution

Ian He

Amateur Explorer of $\mathcal{E}$conϕmetric\$

August 22, 2023

Table of Contents

Overview I

One crucial feature of the variables we study in Econometrics is that the outcome (or payoff) of interest is stochastic, random, and uncertain. Our goal is to study the **distribution** behind such outcome, instead of the probability that a specific outcome appears.

We have multiple approaches to characterizing the distribution of a random variable. Those approaches, in my opinion, can be divided into two big groups:

- Probability-based (direct) approach;
- Expectation-based (indirect) approach.

This time my focus is the second one.

Overview II

	Moment			
Ordinal	**Raw**	**Central**	**Standardized**	**Measuring**
1	Mean			Center
2		Variance		Spread
3			Skewness	Symmetry
4			Kurtosis	Fatness of tail
5+	⋮	⋮	⋮	⋮

Moments

Raw Moment and Central Moment

Definition 1

The mth **(raw) moment** of a random variable X is

$$\mu_m := E(X^m)$$

where E is the expectation operator.

Definition 2

The mth **central moment** of a random variable X is

$$E\left\{[X - E(X)]^m\right\}$$

Important Theorems

The mth moment is said **NOT** to exist when

$$E\left(|X^m|\right) = \int_{-\infty}^{\infty} |x^m| \, dF_X(x) = \infty$$

Theorem 3

If the mth moment about any point exists, so does the $(m-1)$th moment (and thus **all lower-order moments**) about every point.

Theorem 4

For a bounded distribution, the collection of all the moments (of all orders, from 1 to ∞) **uniquely determines** the distribution.

Expectation (1st Raw Moment)

Definition 5

The **expectation** or **expected value** of a discrete random variable X, denoted $E(X)$ or μ_X, is

$$E(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x)$$

where $\mathcal{X}$ is the support set of X and $f_X(\cdot)$ is the probability mass function.

Similarly, the **expectation** or **expected value** of a continuous random variable $X \sim f_X$ is defined as

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) \; dx$$

where $f_X(\cdot)$ is the probability density function.

Properties of Expectation

Some important properties of expectation are

- $E(c) = c$;
- $E(aX + b) = aE(X) + b$;
- $E(c_1 X_1 + c_2 X_2 + \cdots + c_k X_k) = c_1 E(X_1) + c_2 E(X_2) + \cdots + c_k E(c_k)$.

The lowercase letters denote any constants and the uppercase letters denote any random variables.

Alert to Expectation!

1) Expectation is **NOT** everything. Expectation, like any other moment, is just one of the characteristics of a distribution; expectation does **NOT** provide complete information about the distribution.
2) Expectation does **NOT** have to exist. When the support is infinite, the expectation may be infinite; for example, $\sum_{i=1}^{\infty} x_i Pr(X = x_i)$ is a sum of infinitely many terms so it could be infinite. What's worse, when the support of a random variable includes both $-\infty$ and ∞, we even have an *undefined* expectation as the expectation formula produces $\infty + (-\infty) = \infty - \infty$. Anyway, when expectation is infinite or undefined, we say that expectation does not exist.

Variance (2nd Central Moment)

Definition 6

The **variance** of a random variable X is its second central moment:

$$\sigma_X^2 = Var(X) = E\left\{[X - E(X)]^2\right\}$$

It is difficult to interpret the value of variance because it is an average of squared amounts. Instead, we often use the square root of the variance to measure the spread of a distribution:

$$\sigma_X = SD(X) = \sqrt{Var(X)}$$

which is called the **standard deviation**. The standard deviation has the same unit as X.

Properties of Variance

Some important properties of variance are

- $Var(c) = 0$;
- $Var(a + X) = Var(X)$;
- $Var(bX) = b^2 Var(X)$;
- $Var(a + bX) = b^2 Var(X)$; thus, the variance is not a linear operator;
- $Var(X) = E(X^2) - [E(X)]^2$;
- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$, where $Cov(X, Y)$ is the covariance between random variables X and Y.

a, b, and c denote any constants.

Standardized Moments

Definition 7

The mth **standardized moment** (also called the normalized mth central moment) of a random variable X is its mth central moment divided by σ_X^m.

$$\frac{E[(X-\mu_X)^m]}{\sigma_X^m} = \frac{E[(X-\mu_X)^m]}{[Var(X)]^{\frac{m}{2}}}$$

where μ_X is the mean and σ_X is the standard deviation.

For example, skewness is the third standardized moment, and kurtosis is the fourth standardized moment.

Skewness (3rd Standardized Moment) I

Definition 8

The **skewness** of a random variable X is defined as

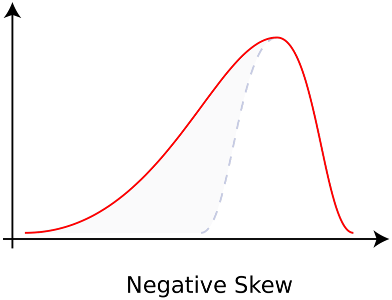
$$Skew(X) = \frac{E[(X - \mu_X)^3]}{\sigma_X^3} = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^3\right]$$

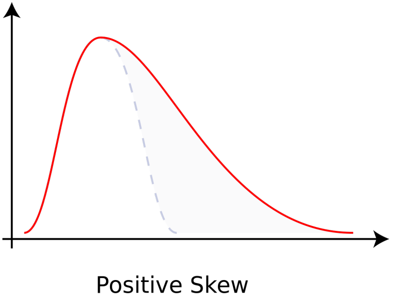
It is sometimes referred to as **Pearson's moment coefficient of skewness** or simply **moment coefficient of skewness**.

Skewness (3rd Standardized Moment) II

The skewness is the measure of the lopsidedness of a distribution.

- A distribution with negative skewness has a longer left tail.
- A distribution with positive skewness has a longer right tail.
- A symmetric distribution has zero skewness.





Source: Wikipedia

Kurtosis (4th Standardized Moment)

Definition 9

The **kurtosis** is defined as

$$Kurt(X) = \frac{E[(X-\mu_X)^4]}{\sigma_X^4} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^4\right]$$

Kurtosis (originating with Karl Pearson) is a measure of the **heaviness of the tail** of the distribution of a random variable. This number is related to the tails of the distribution, **NOT** its peak! Hence, the often-seen characterization of kurtosis as "peakedness" is actually incorrect (e.g., Stata 18 manual, page 2860).

Excess Kurtosis I

Definition 10

The **excess kurtosis** of a random variable X is defined as

$$Kurt(X) - 3$$

The excess kurtosis compares how tail-heavy a distribution is with respect to a **normal distribution** (regardless of its mean and standard deviation). The number 3 in the definition above is the kurtosis of any univariate normal distribution.

Some texts use **Pearson's kurtosis** to refer to the kurtosis (Definition 9) and use **Fisher's kurtosis** to refer to the excess kurtosis (Definition 10).

Excess Kurtosis II

Based on the value of kurtosis, distributions are classified to three groups.

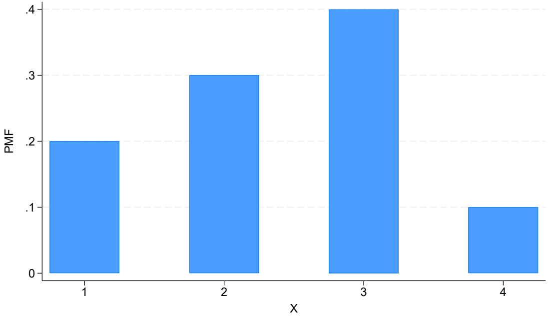
- A distribution is **mesokurtic** (or **mesokurtotic**) if its excess kurtosis is zero.
- A distribution is **leptokurtic** (or **leptokurtotic**) if its excess kurtosis is positive.
- A distribution is **platykurtic** (or **platykurtotic**) if its excess kurtosis is negative.

See appendix if you are interested in the etymology of these strange words.

Numerical Example I

Consider a discrete random variable X with the following PMF:

X	Probability
1	0.2
2	0.3
3	0.4
4	0.1



Q: What are the mean, the variance, the skewness, and the kurtosis of X?

Numerical Example II

The mean of X is

$$E(X) = 1 \times 0.2 + 2 \times 0.3 + 3 \times 0.4 + 4 \times 0.1 = 2.4$$

The variance of X is

$$\begin{aligned} Var(X) = & (1 - 2.4)^2 \times 0.2 + (2 - 2.4)^2 \times 0.3 \\ & + (3 - 2.4)^2 \times 0.4 + (4 - 2.4)^2 \times 0.1 = 0.84 \end{aligned}$$

The standard deviation of X is

$$SD(X) = \sqrt{Var(X)} = \sqrt{0.84} \approx 0.9165$$

Numerical Example III

The skewness of X is

$$
\begin{aligned}
Skew(X) = \frac{1}{0.9165^3}\Big[&(1-2.4)^3 \times 0.2 + (2-2.4)^3 \times 0.3 \\
&+ (3-2.4)^3 \times 0.4 + (4-2.4)^3 \times 0.1\Big] \\
\approx -0.0935 < 0&
\end{aligned}
$$

so X follows a left-skewed distribution.

Numerical Example IV

The kurtosis of X is

$$\begin{aligned} Kurt(X) = \frac{1}{0.9165^4} \Big[& (1-2.4)^4 \times 0.2 + (2-2.4)^4 \times 0.3 \\ & + (3-2.4)^4 \times 0.4 + (4-2.4)^4 \times 0.1 \Big] \\ \approx 2.1020 & \end{aligned}$$

and the excess kurtosis is $Kurt(X) - 3 \approx -0.8980 < 0$, so X follows a platykurtic distribution.

Estimation of Moments

In real life, the population is not (completely) observed; instead, we can only estimate the moments of a distribution by using a sample.

The population mean is estimated by using sample mean (i.e., the arithmetic mean), the population variance is estimated by sample variance, and so forth.

To make the estimation unbiased, we have to do some degree-of-freedom adjustments when estimating variance and higher-order moments. As a result, the calculation becomes unfortunately intense. Plus, be careful that different software packages may employ different adjustments.

An Example

For example, the natural estimator for population variance is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. By contrast, a common corrected-for-bias estimator is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\hat{\sigma}^2$$

Some common adjustments applied to sample skewness and sample kurtosis can be found in their corresponding Wikipedia pages.

MGF and CF

What Is MGF?

There is a useful technical tool which can help us compute every possible moment (if exists). Its name is **moment generating function (MGF)**.

The MGF of a random variable X is

$$M_X(t) = E[\exp(tX)] \quad \text{for } t \geq 0$$

when the expectation exists.

What's the interpretation of t and why does it appear here? Actually, t has no particular interpretation; it's just a device we introduce in order to be able to use calculus. As of now, I haven't found any documents explicitly stating who invented this useful tool.

MGF and Raw Moments

The MGF has the following useful properties:

$$\left.\frac{d}{dt}M_X(t)\right|_{t=0} = E(X)$$

$$\left.\frac{d^2}{dt^2}M_X(t)\right|_{t=0} = E(X^2)$$

$$\vdots \qquad\qquad \vdots$$

$$\left.\frac{d^m}{dt^m}M_X(t)\right|_{t=0} = E(X^m)$$

Characterizing Distributions

The MGF provides a *complete* characterization of a distribution. In other words, each distribution has a *unique* MGF.

Theorem 11

Let X and Y be two random variables, and suppose that the MGFs of X and Y exist and are equal for all $t \in \mathbb{R}$,

$$M_X(t) = M_Y(t)$$

Then, X and Y have the same distribution, and their PMFs or PDFs satisfy

$$f_X(u) = f_Y(u) \quad \forall u$$

Example: MGF of the Standard Uniform Distribution I

Let $X \sim \mathcal{U}[0,1]$. Its PDF is

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Then, its moment generating function is

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \cdot f_X(x)\ dx \\ &= \int_0^1 e^{tx} \cdot 1\ dx \\ &= \left[\frac{1}{t} e^{tx}\right]_0^1 = \frac{e^t - 1}{t} \end{aligned}$$

Example: MGF of the Standard Uniform Distribution II

How to get the mean of this distribution by using the MGF above? Taking the first derivative of $M_X(t)$ with respect to t,

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}\left(\frac{e^t - 1}{t}\right) = \frac{te^t - e^t + 1}{t^2}$$

Taking $t \to 0$, we obtain

$$E(X) = \left.\frac{d}{dt}M_X(t)\right|_{t\to 0} = \lim_{t\to 0}\frac{te^t - e^t + 1}{t^2} = \lim_{t\to 0}\frac{te^t}{2t} = \lim_{t\to 0}\frac{e^t}{2} = \frac{1}{2}$$

where the third equality holds due to the L'Hôpital's rule.

What Is CF?

A major limitation of the MGF is that it does NOT exist for many distributions (e.g., Pareto distribution). Essentially, the existence of the MGF requires the tail of the distribution to decline exponentially; otherwise, the integral (or the sum) would be too large.

This limitation is removed if we use the **characteristic function (CF)**, which is defined as

$$C_X(t) = E[\exp(itX)]$$

where $i = \sqrt{-1}$ is the imaginary unit. The CF always exists because $\exp(itX)$ is always bounded for all $t \in \mathbb{R}$. proof

CF and Moments

The CF has similar properties as the MGF:

$$\frac{d}{dt}C_X(t)\Big|_{t=0} = iE(X)$$

$$\frac{d^2}{dt^2}C_X(t)\Big|_{t=0} = i^2E(X^2) = -E(X^2)$$

$$\vdots \qquad \vdots$$

$$\frac{d^m}{dt^m}C_X(t)\Big|_{t=0} = i^mE(X^m)$$

Similarly, there is a one-to-one correspondence between CF and distribution. Therefore, for any two random variables X and Y, they follow the same distribution if and only if $C_X(t) = C_Y(t)$.

Appendix

Etymology of Kurtosis

Kurtosis was coined by Karl Pearson in about 1895. This word is derived from Ancient Greek *kurtós* ("bulging").

Leptokurtic, **platykurtic**, and **mesokurtic** are built on a combination of different Ancient Greek adjectives, *kurtós*, and an English suffix *-ic*.

- leptokurtic = *leptós* ("thin") + *kurtós* + *-ic*
- platykurtic = *platús* ("flat") + *kurtós* + *-ic*
- mesokurtic = *mésos* ("middle") + *kurtós* + *-ic*

Figure: Platypus

back

Existence of Characteristic Function

To see why the CF $C_X(t)$ exists for all real t, we observe (in the continuous case) its absolute value satisfies

$$|C_X(t)| = \left| \int_{-\infty}^{\infty} e^{itx} f_X(x)\ dx \right| \leq \int_{-\infty}^{\infty} \left| e^{itx} f_X(x) \right|\ dx$$

Since the PDF $f_X(x)$ is non-negative, then $|f_X(x)| = f_X(x)$. In addition, Euler's formula implies

$$\left| e^{itx} \right| = |\cos(tx) + i \cdot \sin(tx)| = \sqrt{\cos^2(tx) + \sin^2(tx)} = 1$$

Thus,

$$|C_X(t)| \leq \int_{-\infty}^{\infty} f_X(x)\ dx = 1$$

which shows that $C_X(t)$ is bounded. Accordingly, $C(t)$ exists for all real values of t. back

References

Blitzstein, J. K. and Hwang, J. (2019).
Introduction to Probability.
CRC Press, 2 edition.

Hansen, B. E. (2022).
Probability and Statistics for Economists.
Princeton University Press.